How task format affects cognitive performance: a memory test with two species of New World monkeys

Schubiger, Michèle N.
Kissling, Alexandra
Burkart, Judith M.

The published article is available from doi:
10.1016/j.anbehav.2016.08.005

How task format affects cognitive performance: a memory test with two species of New World monkeys

Michèle N. Schubiger[1,2], Alexandra Kissling[1] & Judith M. Burkart[1]

[1]Department of Anthropology
University of Zurich
Winterthurerstrasse 190
CH-8057 Zurich
Switzerland

[2]Division of Psychology
School of Social & Health Sciences
Abertay University
Kydd building
Dundee
Scotland

Corresponding author: Michèle N. Schubiger
michele.schubiger@uzh.ch/mnschubiger@gmail.com

Phone: +41 44 635 54 16
Fax: +41 44 635 68 04

Highlights

- Marmosets and squirrel monkeys were tested with two formats of a memory test.
- Performance was strongly affected by task format in both species.
- More options made random choices costly and increased the subjects' motivation.
- This finding has far-reaching consequences for comparisons within & across species.

Abstract

In cognitive tests, animals are often given a choice between two options and obtain a reward if they choose correctly. We investigated whether task format affects subjects' performance in two-choice cognition tests. In experiment 1, a 2-choice memory test, 15 marmosets (*Callithrix jacchus*) had to remember the location of a food reward over time delays of increasing duration. We predicted that their performance would decline with increasing delay, but this was not found. One possible explanation was that the subjects were not sufficiently motivated to choose correctly when presented with only two options because in each trial they had a 50% chance of being rewarded. In experiment 2, we explored this possibility by testing naïve marmosets (n = 8) and squirrel monkeys (*Saimiri sciureus*, n=7) with both the traditional two-choice test *and* a new 9-choice version of the memory test that increased the cost of a wrong choice. We found that task format affected the monkeys' performance. When choosing between nine options, both species performed better and their performance declined as delays became longer. Our results suggest that the 2-choice format compromises the assessment of physical cognition, at least in memory tests with these New World monkeys, whereas providing more options, which decreases the probability of obtaining a reward when making a random guess, improves both performance and measurement validity of memory. Our findings suggest that 2-choice tasks should be used with caution in comparisons within and across species because they are prone to motivational biases.

Keywords: Marmosets, memory, physical cognition, squirrel monkeys, task format

Introduction

When the cognitive abilities of animals are assessed with cognitive tests, subjects are often presented with two options to choose from and rewarded with a food item if they choose the correct option. This two-choice task format has been used to test a variety of cognitive abilities in a range of animal species such as memory (e.g. delayed response tasks in bees, *Apis mellifera;* pigeons, *Columba livia*; several rat strains and many other species, including primates; reviewed in Lind, Enquist, & Ghirlanda, 2015), understanding intentional deception (chimpanzees, *Pan troglodytes*, Woodruff & Premack, 1979; and dogs, *Canis familaris,* Petter, Musolino, Roberts & Cole, 2009) and inferential reasoning (dogs, *Canis familiaris,* Erdöhegyi, Topál, Virányi & Miklósi, 2007*;* birds, carrion crows, *Corvus corone corone,* Mikolatsch, Kotrschal, & Schloegel, 2012; and primates, chimpanzees, *Pan troglodytes*, bonobos, *Pan paniscus*; orangutans *Pongo pygmaeus*; and gorillas *Gorilla gorilla*, Call, 2006). One test that has extensively used the two-choice format in particular with a wide range of animal species is the object-choice task. The object-choice task tests for socio-cognitive abilities by assessing a subject's ability to use an experimenter's gestural cues (e.g. gaze, point, touch) in order to locate a reward that is hidden under one of usually two containers. The range of tested species spans from primates (all four great apes and some Old and New World monkeys), domesticated mammals (dogs, *Canis familiaris*; foxes, *Vulpes vulpes*; cats, *Felis catus;* horses, *Equus caballus*; and goats, *Capra hircus*) and undomesticated terrestrial (wolves, *Canis lupus*; and bats, *Pteropus spp*.) and marine mammals (dolphins, *Tursiops truncates*; seals *Halichoerus grypus* and *Arctocephalus pusillus*; and sea lions, *Otaria byronia*), to corvids (jackdaws, *Corvus monedula;* and nutkrackers, *Nucifraga columbinana*) and parrots (African grey parrot, *Psittacus erithacus);* see Mulcahy & Hedge (2012) for a review.

Although the two-choice task format is widely used in comparative psychology, there is recent evidence that in some circumstances the task may not be a suitable method for assessing

cognitive abilities. Burkart & Heschl (2006), for instance, found that common marmosets (*Callithrix jacchus*), a New World monkey species, chose at random when presented with only two containers in an object-choice task, but they were able to use the experimenter's cues much more reliably and made more correct choices when presented with nine instead of only two containers to choose from. A likely explanation is that lowering the probability of obtaining a reward by random choice helped the marmosets to overcome an inherent social bias that makes non-human primates reluctant to follow communicative cues to food rewards.

In physical cognition tasks, such social biases should not influence a subject's performance, because these tasks usually do not involve any social interaction between subject and experimenter. Memory tests such as delayed response tasks (e.g. Kendrick, Rilling & Denny, 1986; Lind et al., 2015; Rodriguez & Paule, 2009;) for instance, often require the subjects to first observe and later remember in which of two locations a reward has been hidden without obtaining any communicative cues. Consequently, if social biases alone were responsible for the effect of task format on the marmosets' performance in the object choice task, lowering the chance probability of success should not affect their performance in such non-social cognition tasks. Nevertheless, the subjects may prefer to choose in a random manner for other reasons, for instance to avoid the effort of memorizing. To date, it is not known if, or to what extent, task format and chance probabilities also affect performance in physical cognition tests. But if they do so in a similar way, as demonstrated for social tests, this has far-reaching consequences for the validity of species comparisons that are often based on tasks that differ in format.

In the present study, we tested New World monkeys with a physical cognition test that assesses their memory ability and investigated if an alternative task format with nine choices would also be more suitable than the traditional 2-choice task format. In experiment 1, we tested common marmosets *(Callithrix jacchus)* with a traditional 2-choice memory test, i.e. the memory subtest (hidden reward retrieval) of a cognitive test battery designed to assess general

intelligence in non-human primates (Banerjee et al., 2009). In this traditional delayed response memory test, the subjects had to remember the location of a food reward over various time delays. After watching how a food reward was hidden in one of two locations, the subject could no longer see the reward and had to wait until the delay interval had expired before it could choose one of the two locations. New World monkeys, particularly smaller species such as marmosets (Miles, 1956; Miles, 1957a) and squirrel monkeys (French, 1959; Miles, 1957b,), have been shown to perform worse on such delayed response tasks than Old World monkeys (mainly rhesus macaques) and apes (e.g. Fischer & Kitchener, 1965; Harlow, 1932; Miles & Meyer, 1956; reviewed in: Tomasello & Call, 1997). Even though the methodological details are not always comparable, New World monkeys have also been shown to perform equally well (capuchins, *Cebus apella*) or better (spider monkeys, *Ateles geoffroyi*) than Old World monkeys (long-tailed macaques, *Macaca fascicularis*), and even comparable to great apes (Amici, Aureli, and Call, 2010). Moreover, even smaller monkeys usually still perform well above chance, at least with short delays (comparison of apes and monkeys, Fischer & Kitchener, 1965). We therefore expected the marmosets to pass the traditional memory test in experiment 1. Furthermore, in humans, success to remember a specific memory content declines exponentially the more time has elapsed since its acquisition, a phenomenon known as the forgetting curve (Ebbinghaus, 1885, 1913; hereafter Ebbinghaus effect). In experiment 1, we therefore expected that the marmosets' performance would similarly decline with increasing duration of the time delay if this test accurately measured memory performance. Since the marmosets performed relatively poorly in experiment 1 and did not show an Ebbinghaus effect, we conducted experiment 2. Experiment 2 was designed to assess the effect of reducing the chance to obtain a reward when choosing at random. We tested a new sample of marmosets and squirrel monkeys *(Saimiri sciureus)* and compared their performance in a traditional 2-choice versus our newly developed 9-choice version of the memory test.

Experiment 1: A traditional 2-choice memory test

Methods

Subjects

Fifteen common marmosets (*Callithrix jacchus*), 8 males and 7 females participated in this study. All subjects were housed in social groups consisting of two to six individuals at the Primate Station of the Anthropological Institute and Museum of the University of Zurich. Their indoor enclosures had both daylight and artificial light and were composed of one to three components (depending on group size) measuring 1 m (width) x 2 m (depth) x 2 m (height), each of which was equipped with several climbing structures such as natural branches, a sleeping box, an infrared lamp, and a mulch floor. Whenever the weather conditions allowed it, each group had free access to an outdoor enclosure. The marmosets were fed a vitamin and calcium-enriched porridge in the morning, fresh fruit and vegetables at lunchtime, and gum and mealworms in the late afternoon. In addition, they obtained a daily protein-snack in the afternoon such as pieces of cooked egg. Water was available ad libitum from water dispensers. All subjects were tested between their regular feedings and never food deprived during the study. They could enter and leave the test enclosure through semi-transparent plastic tubes that were connected to their home enclosures and were not handled at any time.

Materials and Set-up

Each subject was tested individually in the same compartment (41 cm x 53 cm x 33 cm) of a larger test enclosure, with its group members present in an adjacent enclosure (100 cm x 122 cm x 78 cm) so that the subject could hear and smell, but not see them during testing. The test compartment had a transparent Plexiglas window front containing two rectangular openings (4 cm x 2.5 cm). The test apparatus consisted of two white opaque cylinder-shaped plastic containers (3.0 cm in height and 5.3 cm in diameter) that were attached to a wooden

185 board (33 cm x 33 cm) placed 2 cm from its front, and was placed on the wooden testing table
186 (40 cm x 40 cm) that was level with the test compartment's floor. The test apparatus could be
187 slid in and out of the subject's reach. The two containers were filled with dark-brown bark
188 mulch that corresponded to the flooring substrate in the marmosets' home enclosures. A small
189 yellow piece of locust (*Schistocerca gregaria*) served as a reward in each trial. At the
190 beginning of each trial, the test apparatus was placed just out of the subject's reach and the two
191 containers were each covered with a rectangular mulch piece of approximately the same size as
192 the container.
193
194 *Procedure*
195 The experimenter stood behind the test apparatus, called the subject's name, said "look"
196 while showing it the reward and started a trial as soon as the subject was attentive. She
197 removed the cover of one of the two containers, placed the food reward in the container and
198 again covered it with the mull piece so that the reward was no longer visible and both
199 containers, the baited and the empty one, remained covered. After the delay interval had
200 expired, she slid the board with the containers toward the test compartment's window. The
201 subject could then make a choice by reaching through one of two rectangular openings in the
202 window and removing the mull cover with its hand(s). There were six delay conditions with
203 increasing time delays of 5, 10, 15, 20, 25, and 30 s. Each test session consisted of 10 trials of
204 one delay condition, if possible conducted on the same day, which resulted in a total of 60 trials
205 per subject. The reward's location was counter-balanced in a pseudo-randomized manner so
206 that a locust piece was hidden five times in the left and five times in the right container but
207 never in the same container in more than two consecutive trials. Prior to entering the actual test
208 sessions, each subject went through a pretest phase in which the experimenter followed the
209 same procedure but did not impose a time delay. After the subject reached criterion (≥ 80%
210 correct choices within a single pretest session of 10 trials), it entered the test phase. At the

beginning of each testing day, the subject received one warm-up trial, again without a time delay. Once a subject had finished the six test sessions, it was retested with one full session without a delay. If the subjects had understood the task, we expected their performance in this retest session to be higher than or at least as high as in the test sessions because the retest involved no memory demand. We used two predefined criteria as to when to stop a test session: 1) the subject did not make a choice in three consecutive trials, and 2) the subject was no longer attentive (not looking at the test apparatus but vigilant towards its surroundings instead) to the task, and/or emotionally aroused (emitting vocalisations of discomfort and showing piloerection of the tail; for definitions see Schubiger, Wüstholz, Wunder, & Burkart, 2015), and indicated it wanted to leave the test compartment (climbing to and rattling on the door on top of the test compartment). If the subject met at least one of these criteria, it was allowed to go back to its home enclosure and the session was continued the following day.

Data scoring and analysis

Of the 12 subjects who completed all test sessions, one male subject (Jugo) only completed five trials of the retest and a second male subject (Vito) did not participate in the retest. Three subjects, two males (Kapi and Kantor) and one female (Kitty), did not complete the whole test phase, which resulted in a final total trial number of 756.

All trials were video recorded. The experimenter coded the subjects' choices live using check sheets and checked all trials a second time using the video clips. Five trials (0.7%) had to be excluded from the analysis owing to ambiguous behaviour of the subject or experimenter error. A second rater coded 20 % of the trials from videos. The Kappa statistic was used to determine the reliability between the two raters. Inter-rater reliability was excellent (*Kappa* = .96, $P < 0.000$, N=150).

We ran a General Linear Mixed Model (GLMM) with delay condition as fixed and subject as random factor to determine whether the delay condition significantly affected the

number of correct choices. Furthermore, we conducted one-sample t-tests to determine in which of the six delay conditions the subjects performed above chance levels (more than 50% correct choices) and whether their retest performance was still in the range of the criterion to which they had been trained in the pretest phase.

Results

In the pretest phase, the marmosets reached criterion (≥ 80% correct choices within a single session) within one to 11 sessions of 10 trials each (Mean = 2.93, SD = 2.55, t_{14} = 3.70, P = 0.002). In the test phase, the marmosets chose the correct container across delay conditions in 59% (SD = 8%) of all trials and thus significantly above chance (t_{14} = 4.04, P = 0.001). The GLMM with delay condition as fixed factor and subject as random factor showed that the duration of the delay had a significant effect on the subjects' performance (F(5, 63.77) = 3.31, P = 0.010). We had also predicted that the subjects' performance in the test phase of experiment 1 would decline with increasing length of the time delay, consistent with the Ebbinghaus effect. However, after an initial decline of the number of correct choices that was in line with this prediction, the subjects showed improved performance in the longest two delay conditions (Fig. 1). A one-sample t-test demonstrated that the marmosets performed significantly above chance after delays of 5 s (Mean = 66%, SD = 12%, t_{14} = 5.12, P = 0.000), 10 s (Mean = 59%, SD = 15%, t_{14} = 2.42, P = 0.030), and 25 s (Mean = 68%, SD = 16%, t_{11} = 4.01, P = 0.002), but not after delays of 15 s (Mean = 48%, SD = 17%, t_{13} = -0.34, P = 0.741), 20 s (Mean = 46%, SD = 16%, t_{12} = -0.81, p = 0.432), and 30 s (Mean = 58%, SD = 17%, t_{11} = 1.70, P = 0.117).

In the retest no-delay condition, the marmosets chose the correct container in 66 % of all trials (SD = 16%), which is significantly above chance (t_{10} = 3.46, P = 0.006) and higher than in five of the six test conditions, but differs significantly from the initial 80% criterion in the pretest (Mean = 83%, SD = 5%); t_{10} = 3.33, P = 0.008).

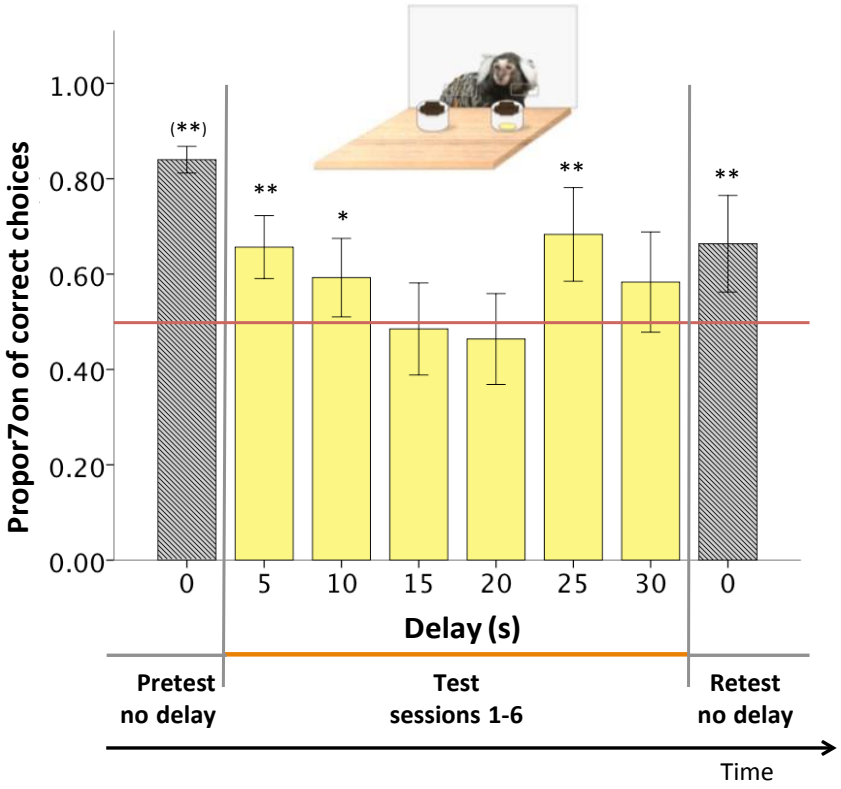

Figure 1 Performance in experiment 1. Subjects had to reach criterion (≥ 80% correct trials within a single session) in the pretest phase (no delay) before entering the test phase (delays = 5 to 30 s) and were retested without a delay after completing the test phase. The red line indicates the chance level of 50%. Significance levels for above chance performance are indicated by *P < .05, **P < .01. *Error bars: 95% confidence intervals.*

Discussion

We tested 15 common marmosets (*Callithrix jacchus)* with a traditional memory test (Banerjee et al, 2009). In this memory test, the subjects had to remember, over several time delay intervals ranging from five to 30 seconds, in which of two locations the experimenter had hidden a reward. All subjects passed the pretest phase, in which no delay was imposed, and, as a group, the marmosets also passed the test phase, by overall performing above chance. In contrast to our predictions, however, the marmosets' performance in the memory test did not decline with increasing delay duration, and they showed quite low levels of correct performance. It is unlikely that the marmosets were unable to remember the reward's location since they performed well after relatively long delays of up to half a minute. Moreover, saddle-back tamarins *(Saguinus fuscicollis),* another callitrichid species and close phylogenetic

relative, have been shown to remember the location of food items over much longer delay intervals of up to 24 hours when tested in a naturalistic foraging task (Menzel, Juno and Garrod, 1985). An alternative explanation for the marmosets' unexpected performance in the hidden reward retrieval test (experiment 1) is that they may not have been sufficiently motivated to choose correctly, particularly after short delays, because of the low cost of a wrong choice. When choosing randomly between the two possible reward locations, they still had a 50% chance of receiving a reward in each trial, and it was only after longer delays between the experimenter's hiding action and the subject's choice that the cost of a wrong choice increased owing to the longer waiting period.

We therefore designed a second experiment to explore if the task format, i.e. the number of choice options, could explain the unexpected pattern of results in the traditional memory subtest. Based on the findings of Burkart & Heschl (2006) in a modified object choice task and our results from experiment 1, we developed a new memory test consisting of nine choice options. This reduced the probability of making a correct choice by chance from 50% in the 2-choice memory test to 11% and thus made a subject's wrong choice more costly. We investigated if this 9-choice format, which had been shown to increase the performance of marmosets in the above mentioned social cognition task, would also be more suitable than the 2-choice format in physical cognition tests. In order to do so, we compared the performance of a naïve marmoset group in the traditional and our new memory test. In addition, we also tested a group of squirrel monkeys (*Saimiri sciureus*), with the same two task formats and directly compared the performance of the two species. This allowed us to evaluate whether task format effects are specific to common marmosets or also present in other non-human primates. We expected both species to perform better in the 9-choice memory test. Furthermore, we expected the squirrel monkeys to outperform the marmosets as in previous delayed response studies (Miles & Meyer, 1956; Miles, 1957b; Treichler, 1964; Tsujimoto & Savaguchi, 2002), owing to their larger absolute and relative brain size (in proportion to their small body size, squirrel

monkeys have the largest brains of all primates; Rowe, 1996), which correlates with general performance in physical cognition tasks (Deaner, van Schaik, & Johnson, 2006; Reader, Hager, & Laland, 2011).

Experiment 2: Introducing a new memory test format

Methods

Subjects

Eight naïve common marmosets *(Callithrix jacchus),* four females and four males, and seven male common squirrel monkeys *(Saimiri sciureus)* who had previous experience with a similar test, participated in this study.

The housing conditions and feeding schedule of the marmosets corresponded to the ones in experiment 1. The squirrel monkeys were housed in two bachelor groups of 3 and 5 individuals, respectively. Their indoor enclosures measured 16.55 m^3 (smaller group) and 24.77 m^3 (larger group) and were equipped with climbing structures, an infrared lamp, and a mulch bark floor. The squirrel monkeys were fed a mixture of pellets and cottage cheese in the morning, a variety of vegetables and a small amount of fruit at lunchtime, and a protein snack such as cockroaches in the late afternoon. Since their indoor enclosures only had artificial UV-light, each group had constant access to a fully roofed outdoor enclosure, and in addition, the two groups took turns in accessing a larger outdoor area of 86.4 m^3 whenever the weather conditions allowed it. The squirrel monkeys could freely travel to and from the test enclosure trough a system of semi-transparent plastic tubes that connected it to their home enclosures.

Set-up

All subjects of both species were tested individually in a separate test compartment of a larger test enclosure. The measurements of the marmosets' test compartment closely resembled the ones in experiment 1, whereas the squirrel monkeys' test compartment measured 110 cm x

98 cm x 77 cm. We again used a test apparatus that could be slid forwards and backwards on a testing table. The apparatuses for the marmosets (M) and the squirrel monkeys (S) were identical and differed only in measurements that were adjusted to the marmosets' smaller body size. It consisted of a wooden frame (M: 40 cm x 37.5 cm/S: 80 cm x 75 cm) containing three wooden platforms (vertical distance between platforms M: 12.5 cm/S: 35 cm) that was mounted on a wooden sliding board (M: 45 cm x 30 cm/S: 95 cm x 50 cm). Empty cylindrical black plastic cups (diameter: 3.1 cm, height: M: 1.1 cm/S: 2.3 cm) with lids were used to hide the reward. For the 9-choice test, three cups where placed equidistant (M: 14 cm/S: 29 cm) between each outer and the middle cup) on each platform (outer cups at M: 4.5 cm/S: 11 cm from the lateral frame). For the 2-choice test, 2 cups were placed on the middle platform (with an in-between distance of M: 11 cm/S: 25 cm and at M: 10 cm/S: 25 cm distance from the lateral frames). In both tests, the cups were held in place by Velcro tape strips. The front of the test enclosure consisted of a lattice that allowed the subjects to reach out and choose one of the cups.

Figure 2 Test apparatus in experiment 2. Shown are both tests: (A) 2-choice, and (B) 9-choice task format. (*Not drawn to scale; note: the lateral parts of the frame were solid boards).*

Procedure

The experimenter's procedure in the pretest and test phase corresponded to the one used in experiment 1 with the exception of two additions in the test phase: 1) The experimenter said

356 "come" while pushing the apparatus toward the subject once the delay had expired in order to
357 encourage the subject to make its choice, and 2) the subject received one to three warm-up
358 trials (no delay) prior to each test session, and the test session only started once it had chosen
359 correctly in a warm-up trial. There were four increasing delay conditions ranging from 5 to 20
360 seconds and each test session consisted of 12 trials of one delay condition. When choosing
361 correctly, the subjects received their favourite rewards, mealworms or cashew nuts (squirrel
362 monkeys) and crickets or cooked apples (marmosets). The same stop criteria as in experiment 1
363 were used to decide when to terminate a session and continue testing on the next day.

364 We used a within-subject design in which every subject of each species was tested with
365 both task formats - the one with two choice options and the one with nine choice options, in
366 counterbalanced order. This resulted in two groups within each species: one group first
367 completed the 2-choice format followed by the 9-choice format while the second group was
368 tested in the opposite order. One male marmoset (Lexus) completed the whole 2-choice
369 memory test but only the 5-s delay condition in the 9-choice memory test. The final sample size
370 therefore consisted of eight marmosets (four females and four males) in the 2-choice and seven
371 marmosets (four females and three males) who completed all conditions in the 9-choice test, as
372 well as seven male squirrel monkeys, who completed both tests.

373

374 **Data scoring and Analysis**

375 All trials were video-recorded and the experimenter coded the subjects' choices live
376 using check sheets. A second rater coded 21 % of the trials from videos. The Kappa statistic
377 was used to determine the reliability between the two raters. For the marmosets one trial had to
378 be excluded owing to experimenter error (no delay imposed). Inter-rater reliability was
379 excellent (100%) for both squirrel monkeys (*Kappa* = 1.00, $P < 0.001$, N=144) and marmosets
380 (*Kappa* = 1.00, $P < 0.001$, N=156).

In order to test which factors best explained the subjects' performance, we ran Generalized Linear Mixed Models (GLMMs) using the Restricted Maximum Likelihood method (REML), with the fixed factors task format, species, delay condition, test order and interactions (species*task format, species*delay, species*order, task format*delay, task format*order, delay*order), and included subject as random factor. The best model was determined using the AICc, the Akaike information criterion corrected for small sample sizes (Hurvich & Tsai, 1989).

Since the probability to be successful by chance differed between the 2-choice and 9-choice format, we could not use the subjects' raw scores to directly compare their performance in the two conditions in the same statistical model but first had to compute a performance measure that was independent of the task format. For this purpose, we computed a performance score for each individual and condition according to the equation below, which corresponds to the square root of the Chi-square value and in which a higher value represents better performance. Observed values correspond to the individual number of correct choices per delay (raw scores of 1 to 12) and expected values were calculated as the number of correct choices expected by chance (6 out of 12 in the 2-choice and 1.33 out of 12 in the 9-choice memory test).

$$\text{Performance score} = \sqrt{\frac{(\text{observed} - \text{expected})}{\text{expected}}}$$

Finally, we conducted one-sample t-tests for each test format to determine in which conditions the subjects performed above chance.

Results

In the pretest phase, the subjects reached criterion (≥ 80% correct within a single session) after one to two sessions (Mean = 1.07; SD = 0.26) in the 2-choice memory test and after one to seven sessions (Mean = 2.27; SD = 1.71) in the 9-choice task. The subjects took significantly longer to reach criterion in the 9-choice than the 2-choice task: $t_{(14)} = -2.61$, $P =$

0.021. There was neither a species-difference in the number of pretest sessions in the 2-choice (squirrel monkeys: Mean = 1.00, SD = 0.00; marmosets: Mean = 1.13, SD = 0.35; t_{13} = - 0.93, P = 0.369) nor in the 9-choice memory test (squirrel monkeys: Mean = 1.57; SD = 1.57; marmosets: Mean = 2.88; SD = 2.10; t_{13} = -1.54, P = 0.015).

The best model included only the fixed effects test format and delay condition and no interactions. Task format had a highly significant effect on the subjects' performance ($F_{1, 98.98}$ = 18.13, P < 0.0001) and so did delay condition ($F_{1, 98.29}$ = 5.65, P = 0.0013). There was no significant effect of species in any of the models (see table 2). Two separate GLMMs based on raw scores of performance, one for each task format, with species, delay and order as fixed factors and subject as random factor demonstrated that delay condition had a significant effect on the subjects' per cent of correct choices for the 9-choice format ($F_{1, 39.88}$ = 5.46, P = 0.003) while there was only a trend for the 2-choice format ($F_{1, 42}$ = 2.49, P = 0.073), see also Figure 3.

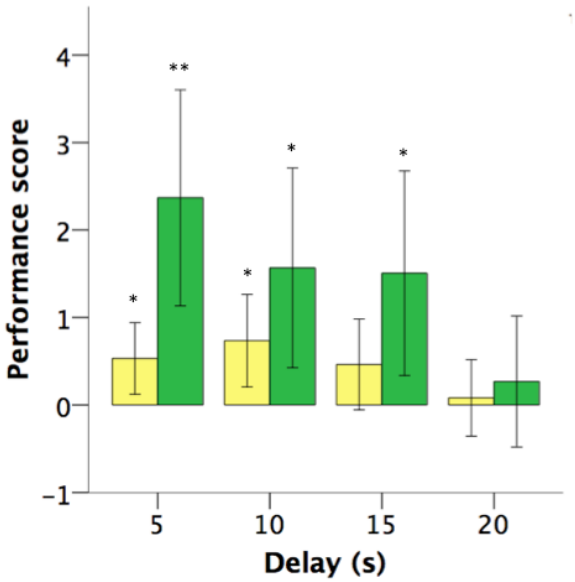

Figure 3 The effect of task format on performance in experiment 2. The subjects' mean performance (χ-transformed test scores to account for the two different chance levels) in the 2-choice (yellow/light bars) and the 9-choice memory test (green/dark bars). *Error bars: 95% confidence interval.* Asterisks * and ** indicate performance significantly above chance (P< .05 and P< .01, respectively) in one-sample-t-tests on the raw values (percent correct choices).

In the 2-choice memory test, mean performance across all test sessions was significantly above chance, i.e. > 50% correct choices (Mean = 59%, SD = 18%, $t_{59} = 4.02$, $P = 0.000$). Split-up per delay condition, the subjects as a group performed significantly above chance in the 5-s (Mean = 61%, SD = 15%, $t_{14} = 2.87$, $P = 0.015$) and 10-s (Mean = 65%, SD = 19%, $t_{14} = 2.97$, $P = 0.010$) delay conditions, but not in the 15-s (Mean = 60%, SD = 19%, $t_{14} = 1.92$, $P = 0.076$) and 20-s conditions (Mean = 52%, SD = 16%, $t_{14} = 0.40$, $P = 0.695$). In the 9-choice memory test, they also performed significantly above chance, i.e. > 11% correct choices, across all test sessions (Mean = 25%, SD = 19%, $t_{56} = 5.43$, $P = 0.000$). Moreover, they performed well above chance after delays of 5 s (Mean = 34%, SD = 21%, $t_{14} = 4.13$, $P = 0.001$), 10 s (Mean = 26%, SD = 19%, $t_{13} = 3.00$, $P = 0.010$), 15 s (Mean = 26%, SD = 20%, $t_{13} = 2.79$, $P = 0.015$), but not 20 s (Mean = 14%, SD = 13%, $t_{13} = 0.786$, $P = 0.446$).

Discussion

In experiment 2, we tested common marmosets and common squirrel monkeys, two evolutionarily closely related species, with both the 2-choice and 9-choice task format of a memory test. As predicted, we found that task format affected the performance of both species. When the subjects were allowed to choose between nine rather than only two options, they performed better, and, in line with our prediction, their performance decreased with longer delays. However, the larger-brained squirrel monkeys did not outperform the marmosets, although the small sample size makes it difficult to identify whether this finding is a true absence of a species difference or instead reflects a lack of statistical power. In contrast to the present findings, squirrel monkeys had outcompeted marmosets in delayed response studies. However, some of these studies did not contain a true memory component (Miles & Meyer, 1956, Miles 1957b) or they contained a stronger working memory component (Tsujimoto & Savaguchi, 2002). In the latter study, the reward's location was not randomized and subjects had to keep in mind their previous choices and base their next choices strategically on these.

Both New World monkey species in our study remembered the reward's location for longer time periods in the 9-choice format than in the 2-choice format. However, their performance was still moderate, for instance in relation to closely related saddle-back tamarins (Menzel et al., 1985) who remembered as many as 30 different locations over much longer delays. Apart from species differences, a likely explanation is that the tamarin study was more naturalistic than ours in that the subjects remained in the group setting during experiments, and the locations to choose from were distributed over a much larger area. In fact, Fischer & Kitchener (1965) had argued that delayed-response tasks with a strong spatial component are easier to solve for non-human primates than non-spatial ones. The more pronounced spatial component may thus have tapped into the tamarins' adaptive behaviour as extractive foragers (Peres, 1992) and thereby made the tasks easier to solve. However, whenever the aim is to assess an individual's cognitive ability per se, rather than a specific adaptation to a narrow cognitive problem, it is preferable to present a task in an abstract rather than a naturalistic manner. This is perhaps most evident for general intelligence test batteries that consist of a number of subtests assessing a wide range of abilities from various cognitive domains to identify whether they are all based on a single domain-general cognitive ability (Burkart, Schubiger & van Schaik, *in press*). The traditional 2-choice memory test in experiment 1 is a subtest of one such test battery, and we developed our 9-choice memory test as a possible alternative.

General discussion

We conducted two experiments in order to explore whether the task format affects cognitive performance of non-human primates in physical cognition tests just as it has been reported for a widely used social cognition test (Burkart & Heschl, 2006). When testing marmosets with a traditional 2-choice memory test (experiment 1), we found that, in contrast to the Ebbinghaus effect, their performance did not continuously decline with increasing delay

duration. To address the possibility that our results reflected a lack of motivation to actually memorize the location of the food rather than the marmosets' ability to do so, we designed a new version of the memory test (experiment 2) with nine choice options instead of two, which lowered the probability of making a correct choice by chance from 50% to 11%. Both marmosets and squirrel monkeys performed better in the 9-choice memory test, and their performance now continuously decreased with increasing delay duration, consistent with the Ebbinghaus effect we had predicted. Our results suggest that the 9-choice format is more accurate in assessing memory performance in the two New World monkey species, and that the 2-choice format negatively affects performance not only in a social cognition task, but also in a physical one.

Our findings have important implications for studies that assess cognitive performance in non-human primates and other animals for comparative purposes. Examples of such comparisons include the assessment of differences in cognitive performance across different tasks between individuals of one species (e.g., to investigate general intelligence; Banerjee et al., 2009; Herrmann, Hernàndez-Lloreda, Call, Hare & Tomasello, 2010), between conspecifics differing in certain traits (e.g., to investigate sex differences; Schubiger et al., 2015) or environmental/ontogenetic conditions (e.g., to investigate rearing differences; Damerius & Forss et al. *in prep*.; Hermann & Call, 2012), and differences in cognitive performance between species (i.e., to investigate evolutionary trajectories; Amici, Aureli & Call, 2008, 2010). For all these comparative purposes it is crucial that differences in measured performance reflect true differences in the subjects' cognitive abilities and cannot be attributed to differences in their motivation to engage with a specific task.

Decreasing the chance-level probability of success, as we have done in the present study, is one way of promoting the subjects' motivation. But although using more than two choice options is advantageous in some cognitive tests with animals, it is probably not feasible in

503 others. Examples for physical cognition tests that require the 2-choice format are the ones in
504 which the subject has to base its choice on more or less apparent differences in the amount (e.g.
505 numerical discrimination tests, Agrillo, 2014), or external features (e.g. tool functionality,
506 Mulcahy & Schubiger, 2014) of the test stimuli. In such tests, additional options could either
507 lead to ambiguous choices or be too demanding for a subject's working memory. However, the
508 costs of a wrong choice can also be increased in 2-choice tests, e.g. by requiring subjects to
509 choose by performing an effortful behavioural response such unscrewing a lid, pulling in the
510 chosen item, or a similarly effortful behaviour.

511 In sum, we found that non-human primates may not be sufficiently motivated to fully
512 engage in a cognitive task when presented in a 2-choice format but that some methodological
513 modifications can restore their motivation.

514 If future studies show that our findings generalize to other species beyond marmosets
515 and squirrel monkeys, and to cognitive domains other than memory, it may be preferable to
516 replace the 2-choice format with alternative task formats. Otherwise, cognitive performance
517 may be biased both in comparisons within and across species, for instance toward more food
518 motivated individuals or species.

519

520 **Acknowledgements**

521 We thank Carel van Schaik and Nick Mulcahy for discussion, Erik Willems for statistical
522 advice, head animal keeper, Heinz Galli, for help with building the experimental apparatuses,
523 and animal keepers, Thomas Bischof and Patricia Rivera, for taking good care of the monkeys.
524 We also thank two anonymous reviewers for their helpful comments. This study was
525 financially supported by the Swiss National Science Foundation (project number 310030-
526 130383) and the A. H. Schultz-Foundation. It was conducted under guidelines established by

the National Veterinary Office of Switzerland and licensed to be conducted by the Veterinary Office of the Canton of Zurich (license number 183/13).

References

Agrillo, C. (2014). Numerical and arithmetic abilities in non-primate species. In Kadosh, R. C., & Dowker, A. (Eds.), .The *Oxford Handbook of Numerical Cognition* (Chapter 12). Oxford University Press: Oxford, UK. DOI:10.1093/oxfordhb/9780199642342.001.0001

Amici, F., Aureli, F., & Call, J. (2008). Fission-fusion dynamics, behavioral flexibility, and inhibitory control in primates. Current Biology, 18(18), 1415-1419. DOI:10.1016/j.cub.2008.08.020

Amici, F., Aureli, F., & Call, J. (2010). Monkeys and apes: are their cognitive skills really so different? American Journal of Physical Anthropology, 143(2), 188-197. DOI:10.1002/ajpa.21305

Banerjee, K., Chabris, C. F., Johnson, V. E., Lee, J. J., Tsao, F., & Hauser, M. D. (2009). General intelligence in another primate: individual differences across cognitive task performance in a New World monkey (*Saguinus oedipus*). *PLoS One*, 4(6), e5883. doi:10.1371/journal.pone.0005883

Burkart, J., & Heschl, A. (2006). Geometrical gaze following in common marmosets (Callithrix jacchus). *Journal of Comparative Psychology*, 120(2), 120-130. DOI:10.1037/0735-7036.120.2.120

Burkart, J. M., Schubiger, M. N., & van Schaik, C. P. The evolution of general intelligence. Behavioral and Brain Sciences. BBS-D-16-00117R1.

Call, J. (2006). Inferences by exclusion in the great apes: the effect of age and species. *Animal Cognition*, 9(4), 393-403. DOI:10.1007/s10071-006-0037-4

Deaner, R. O., Van Schaik, C. P., & Johnson, V. E. (2006). Do some taxa have better domain-general cognition than others? A meta-analysis of nonhuman primate studies. *Evolutionary Psycholo*gy, 4(1), 149-196. DOI: 10.1177/147470490600400114

Ebbinghaus, H. (1885*). Über das Gedächtnis. Untersuchungen zur experiementellen Psychologie*. Duncker & Humber, Leipzig.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.). New York: Columbia University, Teachers College. (Original work published 1885).

Erdöhegyi À., Topál J., Virányi Z., & Miklósi Á. (2007*).* Dog-logic: inferential reasoning in a two-way choice task and its restricted use. *Animal Behaviour* 74, 725–737. DOI: 10.1016/j.anbehav.2007.03.004

Fischer, G. J., & Kitchener, S. L. (1965). Comparative learning in young gorillas and orangutans. *Journal of Genetic Psychology*, 107(2), 337-348.

French, G. M. (1959). Performance of squirrel monkeys on variants of delayed response. *Journal of Comparative and Physiological Psychology*, 52, 741-745.

Harlow H. F. (1932). Comparative behavior of primates. III. Complicated delayed reaction tests on primates. *Journal of Comparative Psychology*, 14:241–252.

Herrmann, E., Hernández-Lloreda, M. V., Call, C., Hare, B., & Tomasello, M. (2010). The structure of individual differences in the cognitive abilities of children and chimpanzees. *Psychological Science*, 21(1), 102-110.

Herrmann, E., & Call, J. (2012). Are there geniuses among the apes? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 367(1603), 2753-2761.

Hurvich, C. M., Tsai, C.-L. (1989). Regression and time series model selection in small samples, *Biometrika*, 76, 297–307.

Kendrick, D. F., Rilling, M. E., & Denny, M. R. (1986). *Theories of animal memory*. Hillsdale, N.J: L. Erlbaum Associates.

Lind, J., Enquist, M., & Ghirlanda, S. (2015). Animal memory: A review of delayed matching-to-sample data. *Behavioural processes*, 117, 52-58.

Menzel, E., Juno, C., & Garrud, P. (1985). Social foraging in marmoset monkeys and the question of intelligence [and discussion]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 308(1135), 145-158.

Mikolasch, S., Kotrschral, K., & Schloegl, C. (2011). Is caching the key to exclusion in

corvids? The case of carrion crows (Corvus corone corone). *Animal Cognition*, 15, 73–82. DOI 10.1007/s10071-011-0434-1

Miles, R. C. (1957a). Delayed-response learning in the marmoset and the macaque. *Journal of Comparative and Physiological Psychology*, 50, 352-355.

Miles, R. C. (1957b). Learning-set formation in the squirrel monkey. *Journal of Comparative and Physiological Psychology*, 50(4), 356-357.

Miles, R. C., & Meyer, D. R. (1956). Learning sets in marmosets. *Journal of Comparative and Physiological Psychology*, 49, 219-222.

Mulcahy, N. J., & Hedge, V. (2012). Are great apes tested with an abject object-choice task? *Animal Behaviour*, 83(2), 313-321. DOI:10.1016/j.anbehav.2011.11.019

Mulcahy, N. J., & Schubiger, M. N. (2014). Can orangutans (*Pongo abelii*) infer tool functionality? *Animal Cognition*, 17(3), 657-669. DOI:10.1007/s10071-013-0697-9

Peres, C. A. (1992). Prey-capture benefits in a mixed-species group of Amazonian tamarins, *Saguinus fuscicollis* and *S. mystax*. *Behavioral Ecology and Sociobiology*, 31, 339–347. DOI: 10.1007/BF00177774.

Petter, M, Musolino, E., Roberts, W. A. & Cole M. (2009). Can dogs (*Canis familiaris*) detect human deception? *Behavioural Processes*, 82, 109–118. DOI:10.1016/j.beproc.2009.07.002

Reader, S. M., Hager, Y., & Laland, K. N. (2011). The evolution of primate general and cultural intelligence. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 366(1567), 1017-1027. DOI:10.1098/rstb.2010.0342

Rodriguez J. S., Paule M. G. (2009). Working Memory Delayed Response Tasks in Monkeys. In Buccafusco J. J, (Ed.). *Methods of Behavior Analysis in Neuroscience* (2 nd ed.). Boca Raton (FL): CRC Press/Taylor & Francis.

Rowe, N. (1996). *The pictorial guide to the living primates*. New York: Pogonias Press.

Schubiger, M. N., Wustholz, F. L., Wunder, A., & Burkart, J. M. (2015). High emotional reactivity toward an experimenter affects participation, but not performance, in cognitive tests with common marmosets (Callithrix jacchus). *Animal Cognition*, 18(3), 701-712. doi:10.1007/s10071-015-0837-5

Tomasello, M., & Call, J. (1997). *Primate Cognition*. New York: Oxford University Press.

Treichler, F (1964). Delayed-response performance by the squirrel monkey. *Psychonomic Science* 1(1), 129-130. DOI:10.3758/BF03342825

Tsujimoto, S., & Sawaguchi, T. (2002). Working memory of action: a comparative study of ability to selecting response based on previous action in New World monkeys (*Saimiri sciureus* and *Callithrix jacchus*). *Behavioural Processes*, 58(3), 149-155.

Woodruff, G., & Premack, D. (1979). Intentional communication in the chimpanzee: The development of deception. *Cognition*, 7, 333-362.